Prof. Philip Koopman

A Safety Case + SPI Metric Approach for Autonomous Vehicle Safety

Carnegie Mellon University

@PhilKoopman

Overview

- Multi-scale metric & feedback loops
  - Design hazard analysis
  - Operational risk mitigation
  - Lifecycle discovery of surprises

https://on.gei.co/2r2rjzg

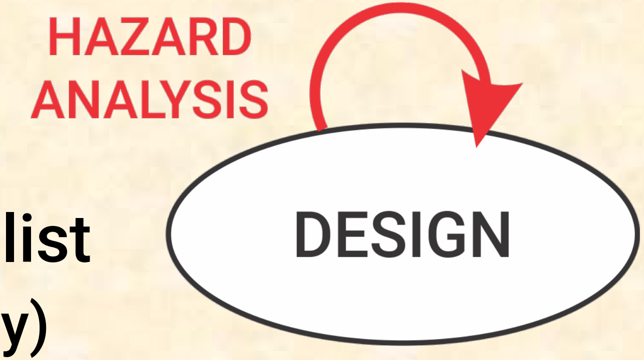
- Safety Performance Indicators (SPIs)
  - Beyond "vehicle acted unsafely"
  - Beyond real-time dynamic risk measurement

  …

  - It's all about monitoring safety case validity

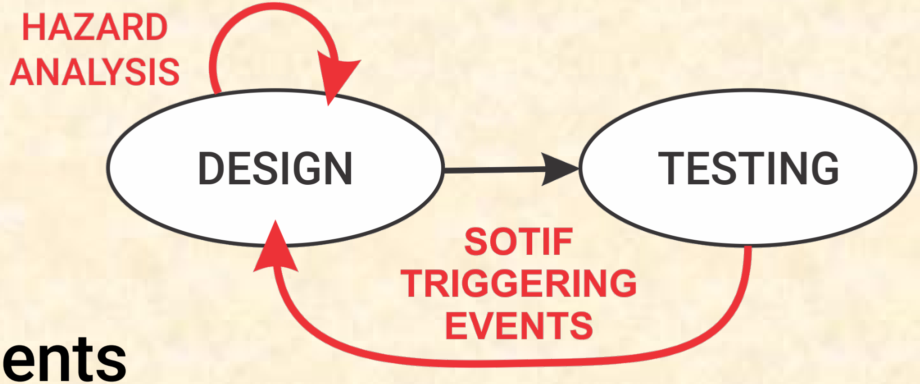
Traditional Hazard Analysis

- Risk Analysis (e.g., start with HARA)
  - List all applicable hazards
  - Characterize the resultant risk
  - Mitigate risk as needed
  - Document all risks acceptably mitigated
- Use various techniques to create hazard list
  - Lessons learned (previous projects; industry)
  - Brainstorming & analysis techniques
    - HAZOP, STPA, …. bring your own favorite approach …
- Limitation: unknown hazards
  - But, human is responsible for overall system safety



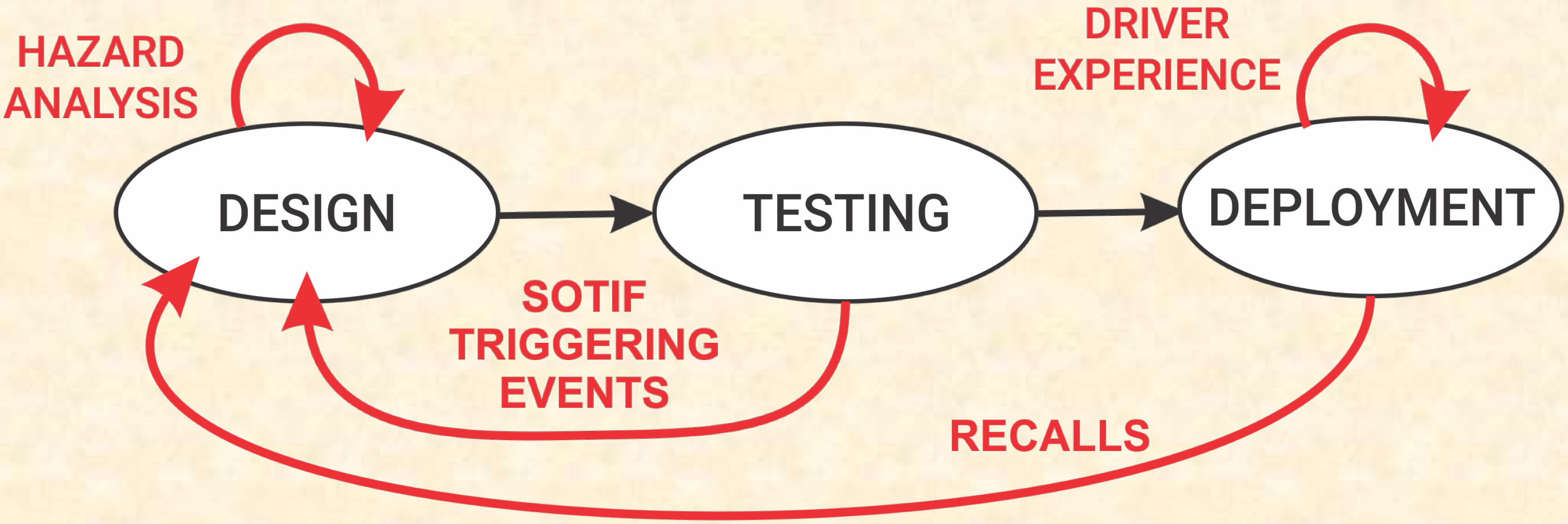
Hazard Analysis for ADAS

- Operating in the open world
  - All hazards aren't known
  - New hazards will appear
- Safety of the Intended Function (SOTIF)
  - Operate in the real world
  - Observe "triggering events"
  - Mitigate discovered hazards
  - Repeat
- Limitation: unseen triggering events
  - But, human is responsible for system safety



Pre-Autonomy & ADAS Feedback Model

- Driver does dynamic risk mitigation
- Recalls for technical faults
  - Recalls are never supposed to happen

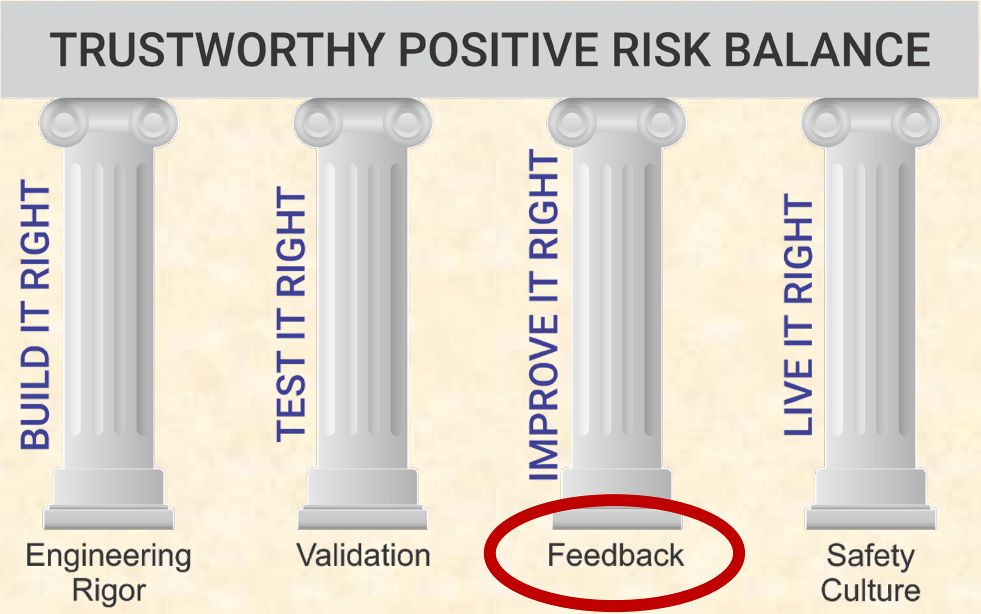


Hazard Analysis for Full Autonomy

- Still an open world with unknowns & changes
  - But … *no human driver responsible*

- Use Positive Trust Balance
  - Engineering rigor
  - Practicable validation
  - Strong safety culture

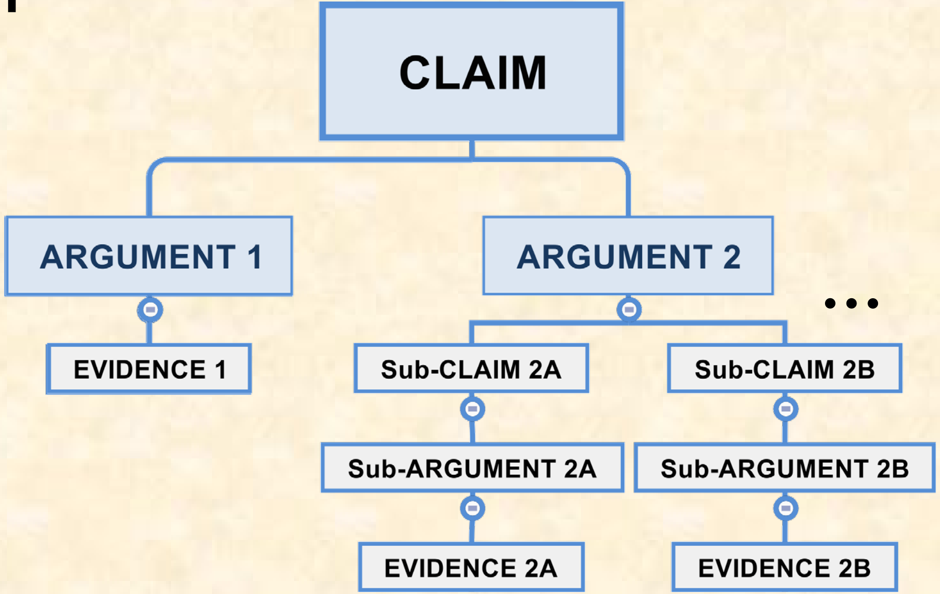
    …. and …
  - Field feedback to handle surprises



- Good fit to UL 4600 ➔ Safety Cases

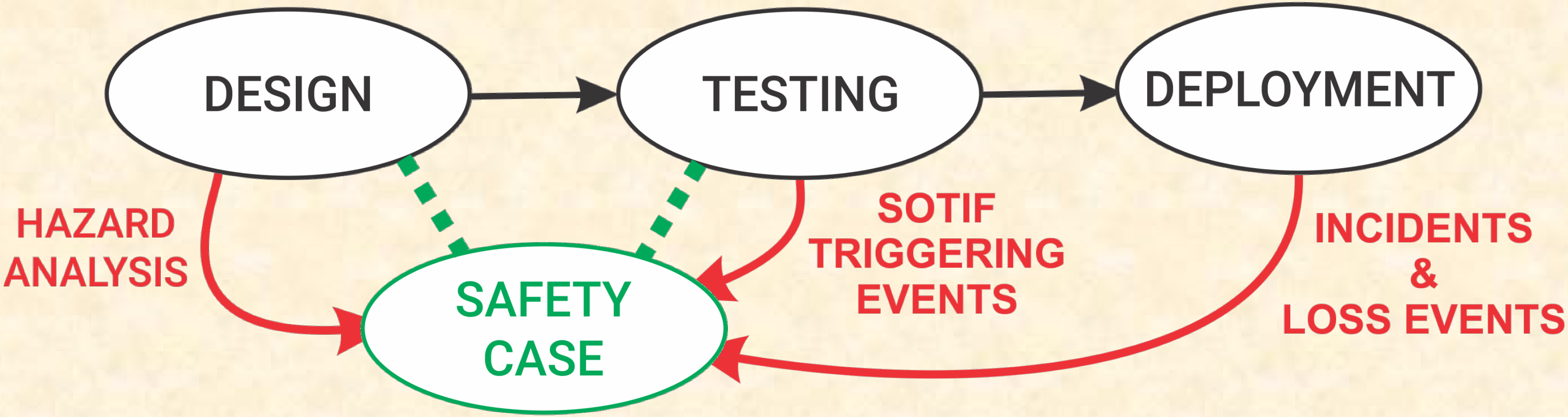
Safety Arguments (Safety Case)

- Claim – a property of the system
  - "System avoids pedestrians"
- Argument – why this is true
  - "Detect & maneuver to avoid"
- Evidence – supports argument
  - Tests, analysis, simulations, …
- Sub-claims/arguments address complexity
  - "Detects pedestrians" // evidence
  - "Maneuvers around detected pedestrians" // evidence
  - "Stops if can't maneuver" // evidence



Default SDC Feedback Model

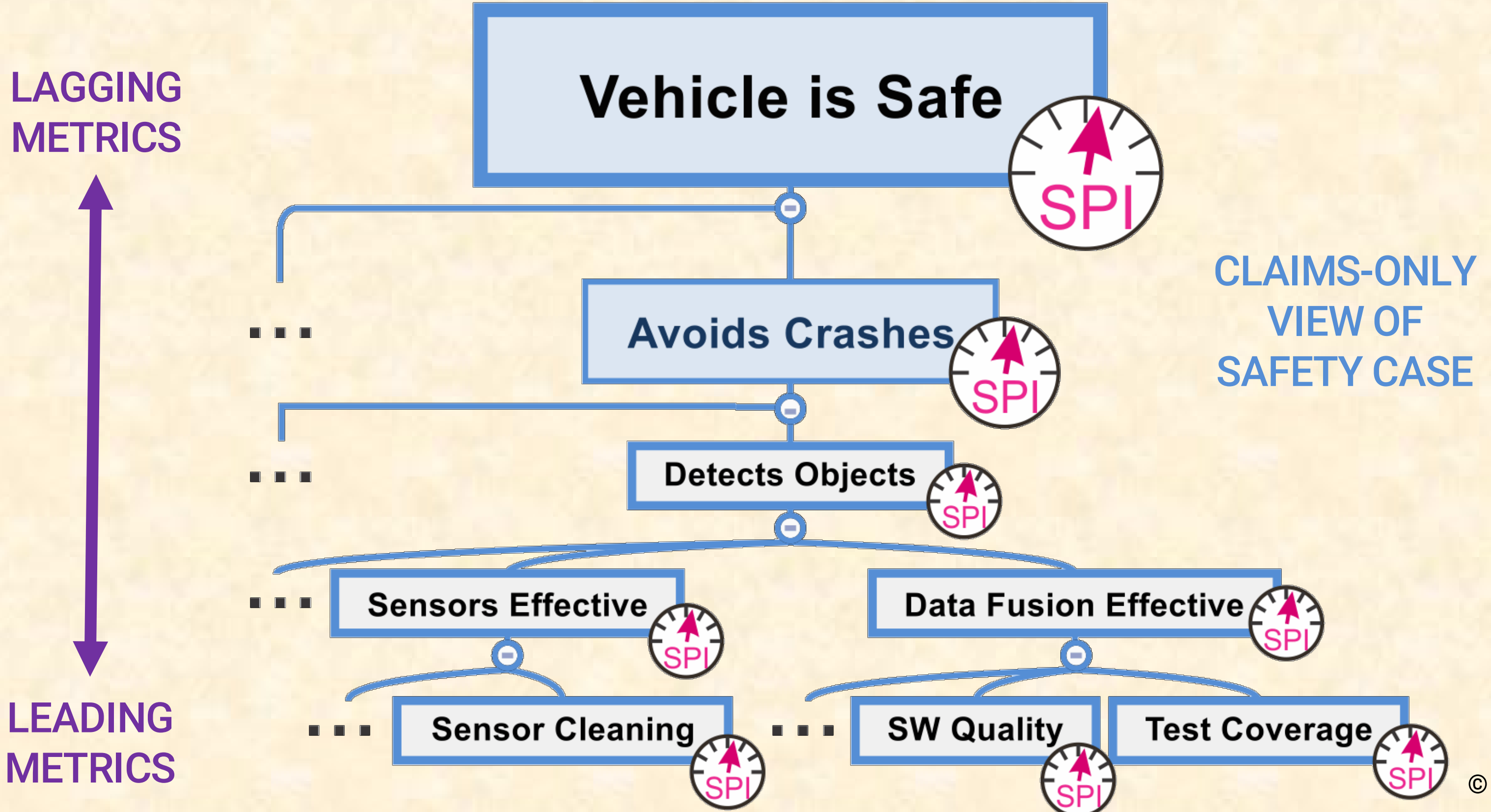
- Safety Case argues acceptable risk – without driver
  - Perhaps Positive Risk Balance (“safer than human”)
  - Update in response to incidents and loss events



  - But, deployment only yields lagging metrics

Safety Performance Indicators (SPIs)

- SPIs monitor the validity of safety case claims



CLAIMS-ONLY VIEW OF SAFETY CASE

Examples of SPIs

- “Acts dangerously” is only one dimension of SPIs
  - Violation rate of pedestrian buffer zones
  - Time spent too close per RSS following distance
- Components meet safety related requirements
  - False negative/positive detection rates
  - Correlated multi-sensor failure rates
- Design & Lifecycle considerations
  - Design process quality defect rates
  - Maintenance & inspection defect rates
- Is it relevant to safety? ➔ Safety Case ➔ SPIs

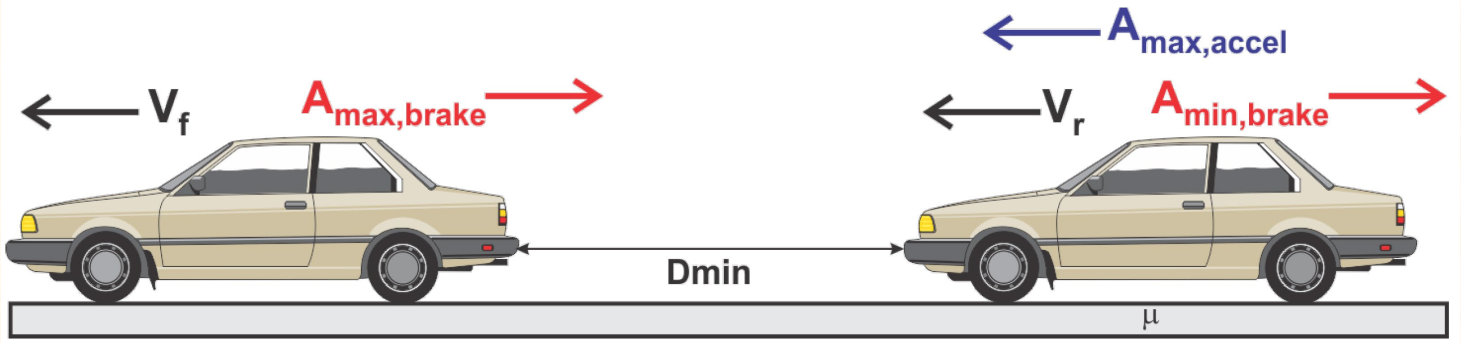
KPI vs. SPI Contrast

stroller_pexels-photo-365813

- Distance to object:
  - KPI: average and variance of clearance
  - SPI: how often SDC violates safe clearance limit
- Sensor effectiveness:
  - KPI: detection rate, SNR per sensor
  - SPI: concurrent multi-sensor detection failure
  - SPI: loss of calibration
- Pedestrian perception:
  - KPI: accuracy, precision, recall
  - SPI: false negative more than <k> consecutive frames
  - SPI: systematic under-performance on sub-classes

Runtime Monitoring Implications

- Responsibility-Sensitive Safety (RSS) Scenario:



- Safety monitor: increase distance if too close in case of panic stop
- KPI: best effort separation given driving conditions
- SPIs: situation more dangerous than expected (e.g., ODD issues)
  - Spent more time in too-dense traffic than expected
  - Lead/own vehicle brake violate expectations
  - Other vehicles panic brake more often than assumed

SPIs and Lifecycle Feedback

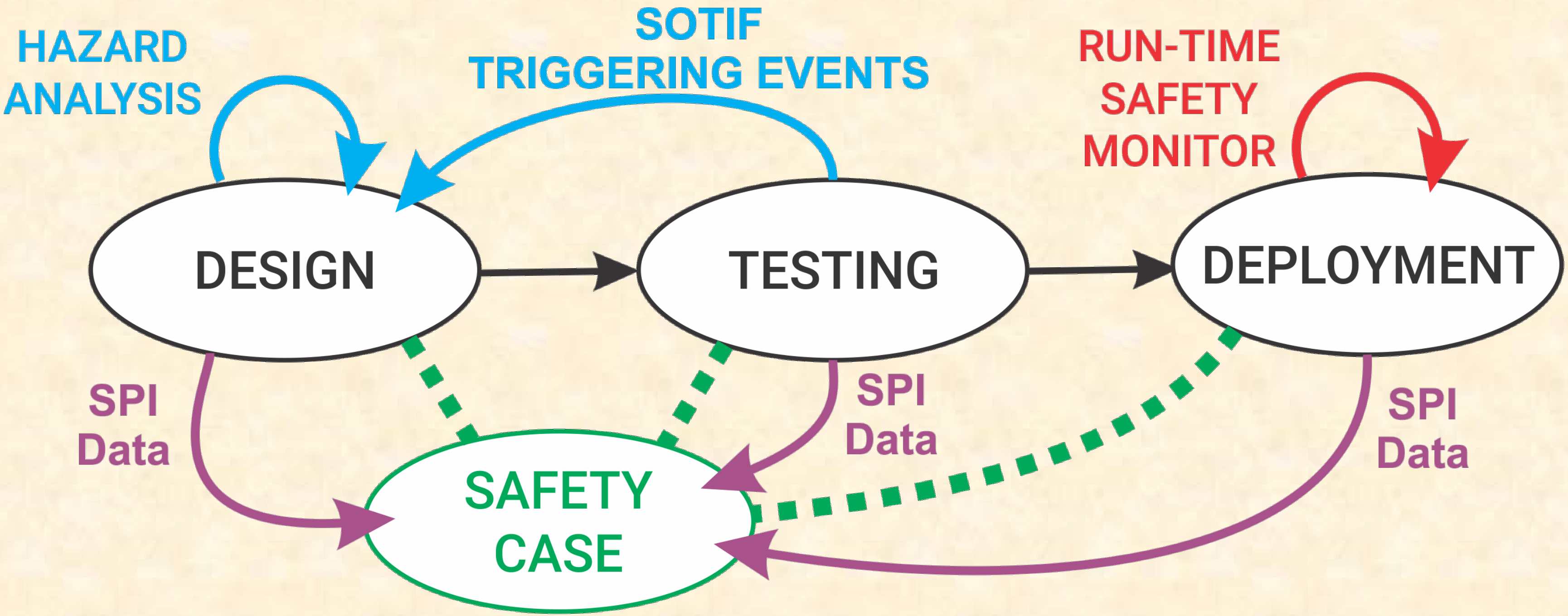
- SPI measures validity of a safety case claim

➔ a SPI value violation means safety case is invalid

- Root cause analysis might reveal:
  - Design process execution defect
  - Design defect
  - Hazard analysis gap
  - SOTIF analysis gap
  - Training data bias
  - Evidence gap, or defect
  - Assumption error

SPI-Based Feedback Approach

- Safety Case argues acceptable risk
  - SPIs monitor validity of safety case



Summary

- Monitoring incidents is only part of feedback
- Removing human means mitigating surprise
  - Tactical: run-time safety monitoring
  - Strategic: run-time SPI monitoring
- SPIs provide feedback on:
  - Design quality & process maturity
  - Testing coverage
  - Lifecycle procedure execution
- SPIs: you are as safe as you think you are
  - Field feedback is key to SPI success

pexels-dom-j-297927.jpg